LINEAR STATISTICAL MODELS

PROJECT – THE UK INTERNET USER ANALYSIS

NAME: BHARATHPRASAD SIVAPRASAD

STUDENT ID: BS330

ABSTRACT

The project analyses the internet usage in the UK by considering the age, sex, disability, ethnicity, and economic activity. The objective of this project is to find the feasibility and scope of the company to start the telecommunications operations in the United Kingdom by analysing the trend of various factors in consideration. R language and Excel is used to create the model. The program codes have been attached in appendix section.

INTRODUCTION

As the United Kingdom market for telecommunication sector is saturated, it is essential to find the scope of the market and the profitability of the business by analysing the data before proceeding with the business plan. In this project, we will analyse and predict for the diverse data provided. The data provided is for the last 3 months of the years from 2014 to 2021. We will analyse the number of internet users with respect to age, sex, disability, ethnicity and economic activity. Throughout the project, different mathematical and statical techniques such as Simple Linear regression and Multiple Linear regression will be used to calculate and predict the data.

Linear Regression is the model which finds the best fit linear regression line between the independent (x) and dependent variable (y). It finds the linear relationship between the dependent and independent variable and models through a random disturbance term (or, error variable) ε. The disturbance is primarily important because we are not able to capture every possible influential factor on the dependent variable of the model. To capture all the other factors, not included as independent variable, that affect the dependent variable, the disturbance term is added to the linear regression model.

$$
\begin{aligned}
y &= \mathbf{X}\beta + \varepsilon \\
&= \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_K \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \varepsilon \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon
\end{aligned}
$$

Linear Regression is of two types: Simple and Multiple. Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable. Whereas, In Multiple Linear Regression there are more than one independent variables.

Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is the difference between the actual value and Predicted value.

The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

To understand the performance of the Regression model performing model evaluation is necessary. Some of the Evaluation metrics used for Regression analysis are:

1. R squared or Coefficient of Determination: The most used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.
2. Adjusted R squared: It is the improvement to R squared. The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So, the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model and shows the real improvement of the model. Adjusted R2 is always lower than R2.
3. Mean Squared Error (MSE): Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.
4. Root Mean Squared Error (RMSE): It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE does not.

ANALYSIS OF INTERNET AND INTERNET NON-USERS WITH RESPECT TO AGE GROUP

For analysing the internet user data, we have considered the age groups of 16-24, 25-34, 35-44, 45-54, 55-64, 65-74 and 75+ for the years 2014 to 2021.

Simple linear regression is applied to model each age groups and it has been analysed separately before applying the multiple linear regression to analyse the age group data for a compiled conclusive analysis.

Age Group 16-24:

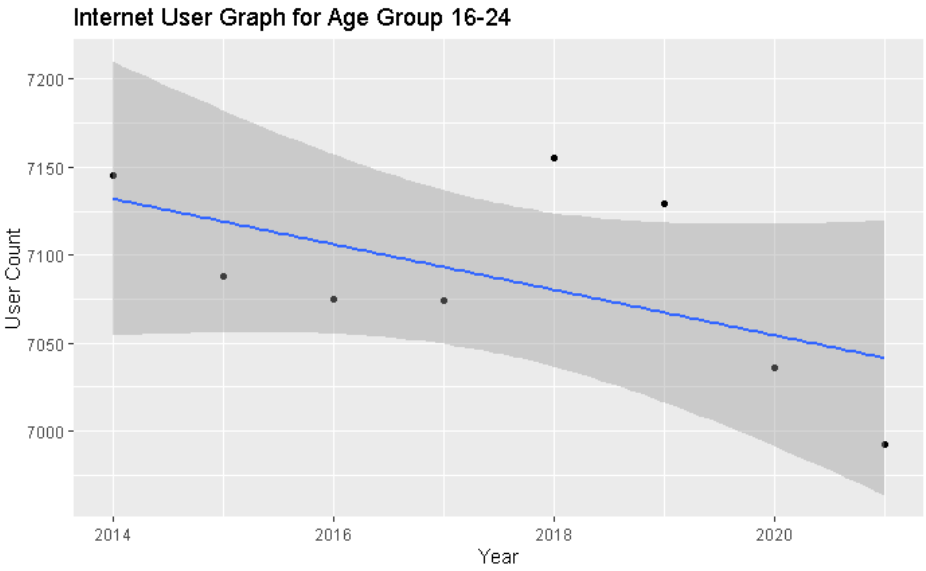

Fig 1.1 Internet User linear regression model – Age (16-24)

Figure 1.1 represents the internet usage linear regression model graph for the age group 16 to 24.

As we can observe from the graph, the linear regression is not following closely with the original data points. And the variation from the regression line with the data points are larger.

While interpreting the parameters from the mathematical calculations, we observed that the p-value is 0.1402 which is greater than 0.05. Based on the p-value we can conclude that the predictor variable (in this case age group 16-24) is insignificant because the larger the 'p' value the more insignificant is the variable and vice versa.

The relationship between predictor and response variable can be answered by observing the F stats. This defines the collective effect of all predictor variables on the response variable. In this model, F=2.887 is greater than 1, but not significantly greater. So it can be concluded that there is a relationship between predictor and response variables but the relationship cannot be concluded as significant.

To conclude whether the model is fit for the predicted and observed data, we should consider the R2 (multiple-R-squared) value as it indicates how much variation is captured by the model. R2 closer to 1 indicates that the model explains the large value of the variance of the model, and it can be considered a good fit. In this case, the value is 0.2123 (distant to 1) and hence the model is not a good fit.

So, we can conclude that mathematically, simple linear regression model is not a fit for given internet users data for the age group 16-24. But by analysing the trend of data points and prediction line we can approximately presume that the internet usage among the age group 16-24 is decreasing year by year.

Age Group 25-34:

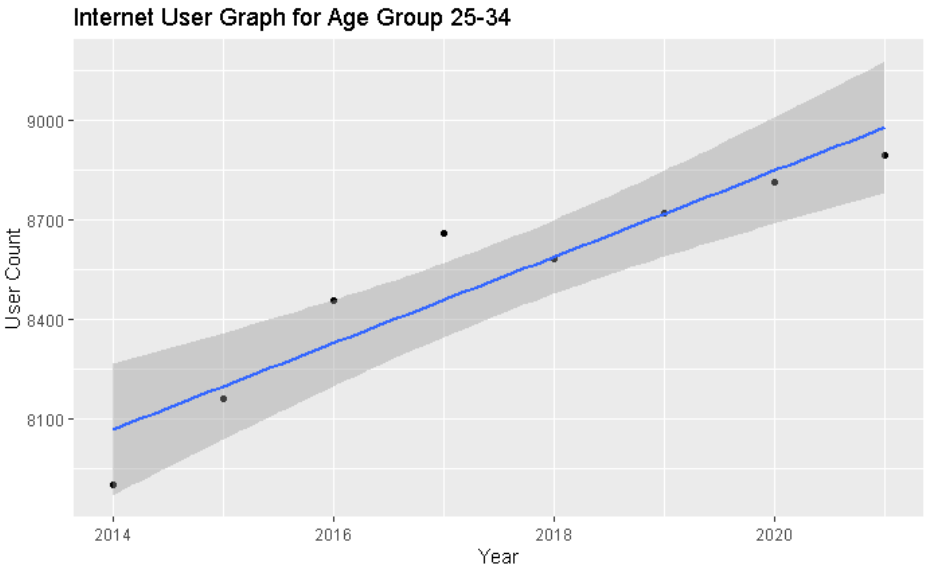

Fig 1.2 Internet User linear regression model – Age (25-34)

Figure 1.2 represents the internet usage linear regression model graph for the age group 25 to 34.

As we can observe from the graph, the linear regression is following closely with the original data points. While interpreting the parameters from the mathematical calculations, we observed that the p- value 0.0005273 which is significantly lesser than 0.05. Based on the p-value we can conclude that the predictor variable (in this case age group 25-34) is significant.

We observed that there is a positive beta estimate (β = 129.92) which indicates that with increase in years, the count of internet users also increases.

Now let us find out the relationship between predictor and response variables. In this model, F= 45.18 is significantly greater than 1. So, it can be concluded that there is a relationship between predictor and response variables.

By calculating the R2 (multiple R-squared) value, let us predict whether the model is fit or not. We observed the R2 value to be 0.8828 which is closer to 1. So, we can conclude that the model is fit the original data.

Conclusion:

Both mathematically and visually, the simple linear regression model is a good fit for the observed data for the internet user data for age group 25-34. The internet usage among this age group tend to increase linearly over the succeeding years.

Prediction for 5 years:

```
       1        2        3        4        5
9108.750 9238.667 9368.583 9498.500 9628.417
```

The prediction for 5 years from 2022 to 2026 is represented above where ‘1’ represents the year to ‘5’ representing the year 2026. The values represent the internet user count for each year.

Age Group 35-44:

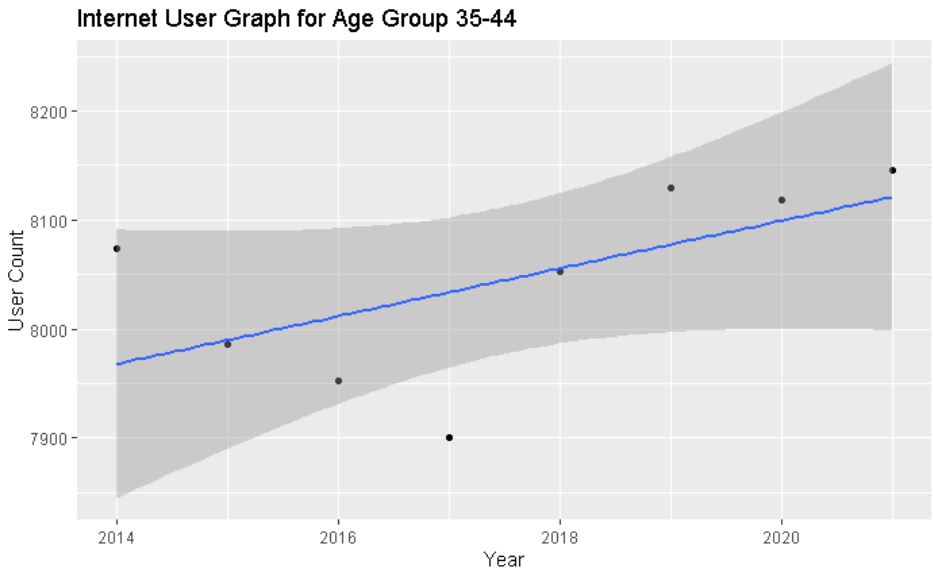

Fig 1.3 Internet User linear regression model – Age (35-44)

Figure 1.3 represents the internet usage linear regression model graph for the age group 35 to 44.

As we can observe from the graph, the linear regression line is following closely with the original data points. While interpreting the parameters from the mathematical calculations, we observed that the p- value 0.1175 which is significantly higher than 0.05. Based on the p-value we can conclude that the predictor variable (in this case age group 35-44) is insignificant.

Now let us find out the relationship between predictor and response variables. In this model, F= 3.337 is greater than 1, but not significant.

By calculating the R2 (multiple R-squared) value, let us predict whether the model is fit or not. We observed the R2 value to be 0.3574 which is distant from the optimal value 1. So, we can conclude that the model is not a good fit with the original data.

Conclusion:

So, we can conclude that mathematically, simple linear regression model is not a fit for the given internet users data for the age group 35-44. But by analysing the trend of data points and prediction line we can approximately presume that the internet usage among the age group 35-44 is increasing year by year.

To avoid repetition of words, let us consider the remaining age groups together and analyse jointly.

Age groups 45-54, 55-64, 65-74 and 75+:

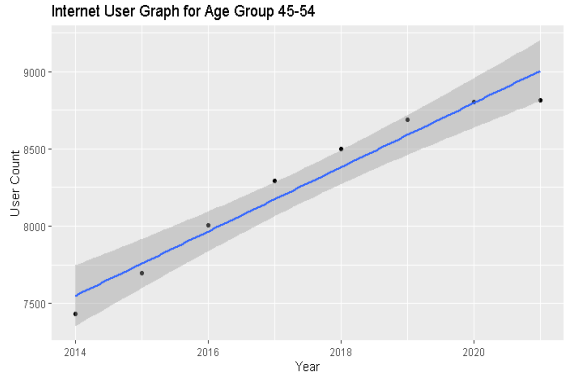

Fig 1.4 Internet User linear regression model – Age (45-54)

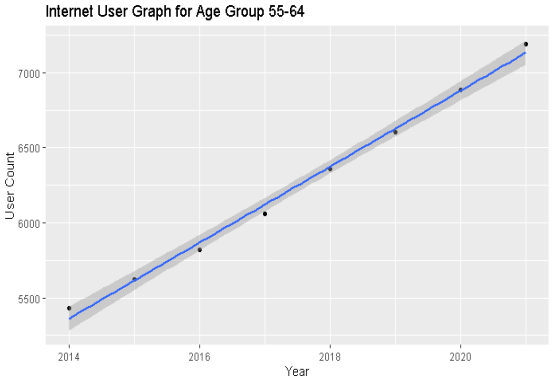

Fig 1.5 Internet User linear regression model – Age (55-64)

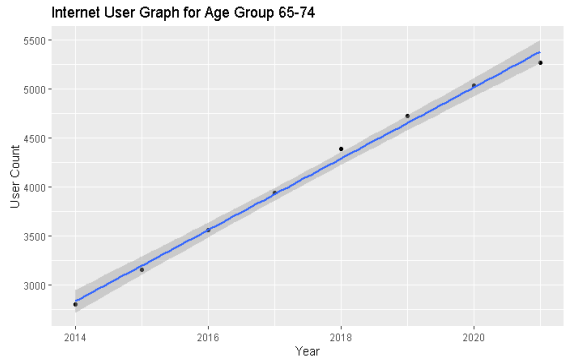

Fig 1.6 Internet User linear regression model – Age (65-74)

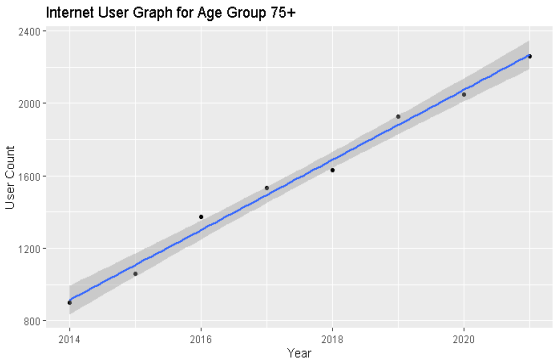

Fig 1.7 Internet User linear regression model – Age (75+)

Figure 1.4, 1.5, 1.6 and 1.6 represents the internet usage linear regression model graph for the age group 45 to 54, 55 to 64, 65 to 74 and above 75 years respectively.

All four plots have a positive linear regression line against the original internet user data with respect to the age groups.

Now, let us examine the p-values associated with each calculation. The p-value observed for the age group 45 to 54 is 3.664e-05, 55 to 64 is 5.29e-08, 65 to 74 is 5.856e-08 and 75+ is 2.606e-07. All the p-values are significantly lower than 0.05. Therefore, we can conclude that the predictor variable is significant for all the considered age groups in this section.

Now let us find out the relationship between predictor and response variables. F-stat value for the age group 45-54 is 117.4, 55-64 is 1079, 65-74 is 1043 and 75+ is 632.2. All the calculated F-stat values are significantly greater than 1. So, it can be concluded that there is a relationship between predictor and response variables for all these age group for internet users.

By calculating the R2 (multiple R-squared) value, let us predict whether the model is fit or not. The R2 for the age group 45-54 is 0.9514, 55-64 is 0.9945, 65-74 is 0.9943 and 75+ is 0.9906. The observed R2 values are closer to 1. So, we can conclude that the models are fit to the original data.

Also, we observed that there is a positive beta estimate (β) for all the age groups which indicates that with increase in years, the count of internet users also increases.

β for age group 45-54 = 208.14

β for age group 55-64 = 253.1

β for age group 65-74 = 363.96

β for age group 75+ = 193.7

In conclusion, we can say that, both mathematically and visually, the simple linear regression model is a good fit for the observed data for the internet user data for age groups 45-54, 55-64, 65-74, and 75+. The internet usage among these age groups tend to increase linearly over the succeeding years.

The summary of all the observed parameters for the linear regression model for age group is given in Table 1.

Age Group	P – Value	R2	F-Statistic	Beta estimate (β)
16 – 24	0.1402	0.3248	2.887	-12.952
25 – 34	0.0005273	0.8828	45.18	129.92
35 – 44	0.1175	0.3574	3.337	21.92
45 – 54	3.664×10^{-5}	0.9514	117.4	208.14
55 – 64	5.29×10^{-8}	0.9945	1079	253.10
65 – 74	5.856×10^{-8}	0.9943	1043	363.96
75+	2.606×10^{-7}	0.9906	632.2	193.70

Table 1 – Summary of parameters of Linear Regression model with respect to Age Group

Prediction for 5 years:

- Age group 45-54:

1	2	3	4	5
9214.143	9422.286	9630.429	9838.571	10046.714

The prediction for 5 years from 2022 to 2026 is represented above where '1' represents the year to '5' representing the year 2026. The values represent the internet user count for each year.

- Age group 55-64:

1	2	3	4	5
7387.143	7640.286	7893.429	8146.571	8399.714

- Age group 65-74:

1	2	3	4	5
5745.214	6109.179	6473.143	6837.107	7201.071

- Age group 75+:

```
       1        2        3        4        5
2462.893 2656.619 2850.345 3044.071 3237.798
```

Analysis of Age group data using Multiple Linear Regression:

```
Residuals:
    Min      1Q  Median      3Q     Max
-729.45 -165.30  -19.42  156.71  637.18

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -326624.22   36596.12  -8.925 9.14e-12 ***
Year            165.41      18.14   9.119 4.75e-12 ***
age_grp25      1437.38     155.51   9.243 3.12e-12 ***
age_grp35       957.88     155.51   6.160 1.43e-07 ***
age_grp45      1190.75     155.51   7.657 7.29e-10 ***
age_grp55      -838.75     155.51  -5.393 2.09e-06 ***
age_grp65     -2979.37     155.51 -19.159  < 2e-16 ***
age_grp75     -5495.62     155.51 -35.339  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 311 on 48 degrees of freedom
Multiple R-squared:  0.9859,    Adjusted R-squared:  0.9838
F-statistic: 478.5 on 7 and 48 DF,  p-value: < 2.2e-16
```

Fig1.8 Multiple linear regression parameters – Age group

Figure 1.8 represents the parameters calculated after executing the given set of data with age group to the multiple linear regression.

The p-value is approximately 2.2×10^{-16} which is significantly lower than 0.05 value. Apart from the p- value, R2 value is approximately 0.9859 which is closer to the optimal value 1. F – statistic is 478.5 which is significantly higher than 1. All these parameter values shows that the given data fits with multiple linear regression model when we consider the age groups as an independent variable.

ANALYSIS OF INTERNET AND INTERNET NON-USERS WITH RESPECT TO ETHNIC GROUP:

In this section, we analyse the effect of ethnicity on the usage of internet in the United Kingdom. The ethnic groups under consideration are White, Mixed/multiple ethnic background, Indian, Pakistani, Bangladeshi, Chinese, Other Asian background, Black/African/Caribbean/Black British and Other ethnic groups.

Simple linear regression is applied to model each ethnic groups with respect to their internet user count and it has been analysed separately before applying the multiple linear regression to analyse the ethnic group data for a compiled conclusive analysis.

Table 2 represents the parameters of linear regression model with respect to the internet user data with different ethnic groups over the years 2014 to 2021.

Age Group	P – Value	R2	F-Statistic	Beta estimate (β)
White	2.485e-07	0.9907	642.4	9.041e+02
Mixed/multiple ethnic background	0.002015	0.8184	27.04	28.226
Indian	0.0007568	0.868	39.47	34.440
Pakistani	4.83e-06	0.9752	235.6	43.869
Bangladeshi	8.078e-05	0.9368	88.96	20.988
Chinese	0.01808	0.6338	10.39	8.679
Other Asian background	0.001912	0.8215	27.61	22.143
Black/African/Caribbean/Black British	7.745e-05	0.9377	90.29	6.592e+01
Other ethnic group	0.001157	0.8484	33.57	29.893

Table 2 - Summary of parameters of Linear Regression model with respect to Ethnic Groups

Now let us plot the linear regression model graph for each ethnic group for internet user count against years.

White:

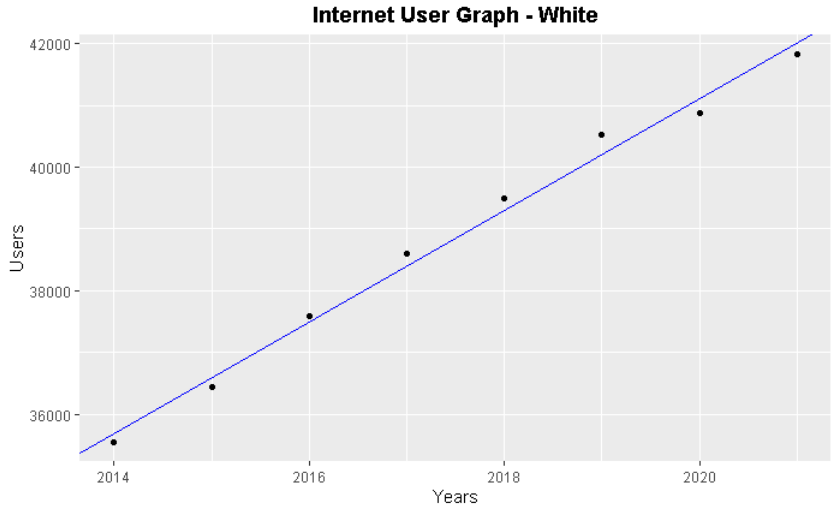

Fig 2.1 - Internet User linear regression model Ethnicity – White

Figure 2.1 represents the internet usage linear regression model graph for the ethnic group White.

As we can observe from the graph, the linear regression is closely following the original data points. While interpreting the parameters from the mathematical calculations, we observed that the p-value 2.485×10^{-07} which is significantly lesser than 0.05. Based on the p-value we can conclude that the predictor variable (in this case ethnicity - White) is significant.

We observed that there is a positive beta estimate ($\beta = 904.1$) which indicates that with increase in years, the count of internet users also increases.

Now let us find out the relationship between predictor and response variables. In this model, F= 642.4 is significantly greater than 1. So, it can be concluded that there is a relationship between predictor and response variables.

By calculating the R2 (multiple R-squared) value, let us predict whether the model is fit or not. We observed the R2 value to be 0.9907 which is closer to 1. So, we can conclude that the model is fit to the original data.

Conclusion:

Both mathematically and visually, the simple linear regression model is a good fit for the observed data, for the internet user data, for the White ethnic group. The internet usage among this ethnicity tend to increase linearly over the succeeding years.

Let us summarize, the analysis for all the other group by analysing the plots and taking Table 2 in consideration, in the following section.

Prediction for the next 5 years:

1	2	3	4	5
42930.64	43834.79	44738.93	45643.07	46547.21

The prediction for 5 years from 2022 to 2026 is represented above where '1' represents the year to '5' representing the year 2026. The values represent the internet user count for each year.

Analysis for all the other Ethnic groups:

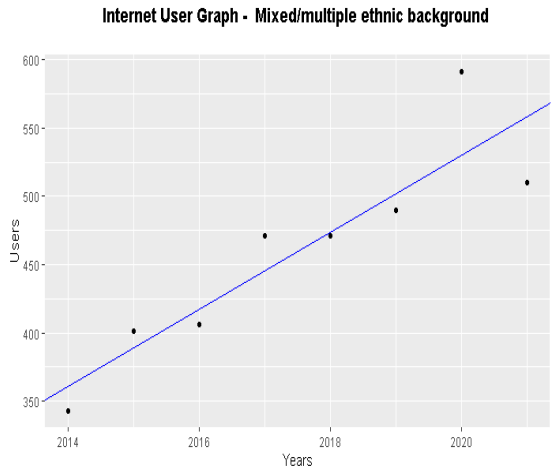

Fig 2.2 - Internet User linear regression model Ethnicity – Mixed/multiple ethnic background

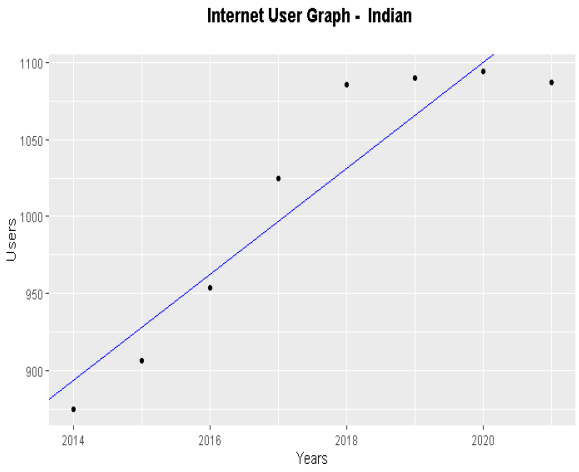

Fig 2.3 - Internet User linear regression model Ethnicity – Indian

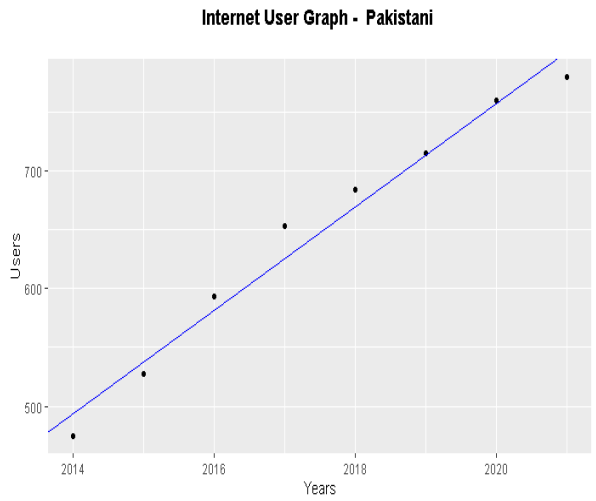

Fig 2.4 - Internet User linear regression model Ethnicity – Pakistani

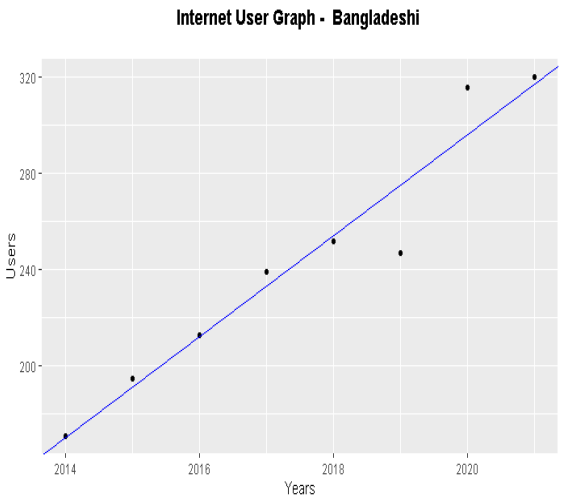

Fig 2.5 - Internet User linear regression model Ethnicity – Bangladeshi

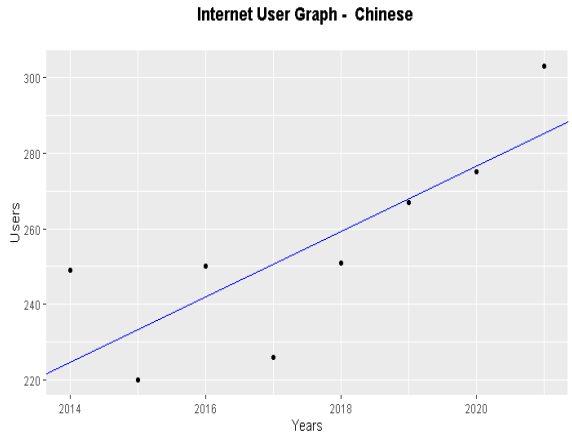

Fig 2.6 - Internet User linear regression model Ethnicity – Chinese

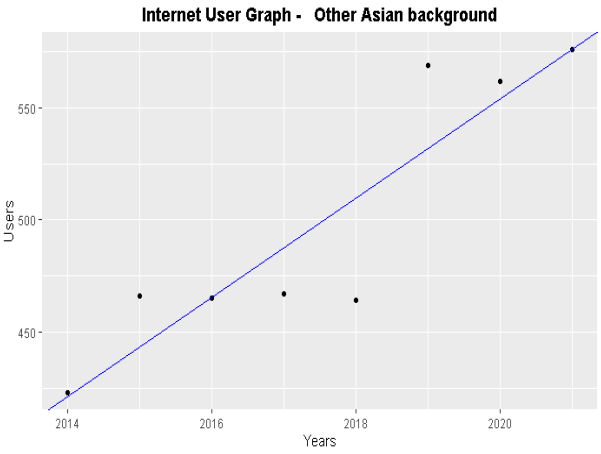

Fig 2.7 - Internet User linear regression model Ethnicity – Other Asian background

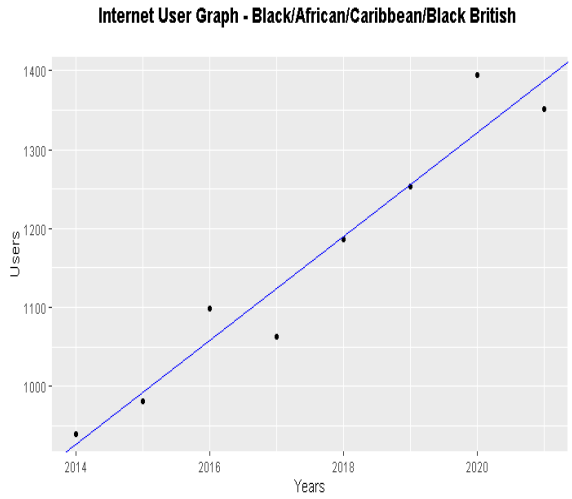

Fig 2.8 - Internet User linear regression model Ethnicity – Black/African/Caribbean/Black British

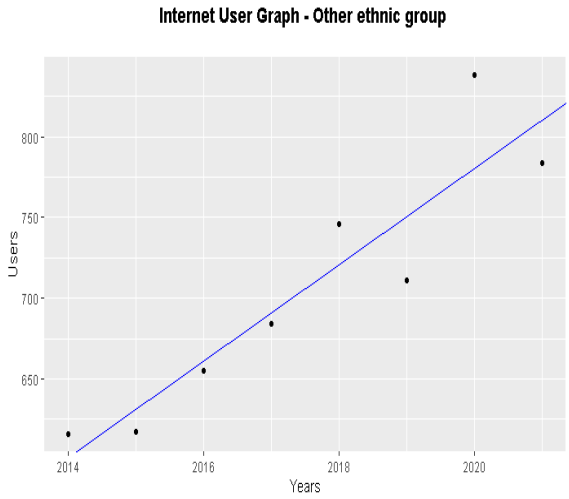

Fig 2.9 - Internet User linear regression model Ethnicity – Other ethnic group

Figure 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8 and 2.9 represents the internet usage linear regression model graph for the ethnic groups Mixed/multiple ethnic background, Indian, Pakistani, Bangladeshi, Chinese, Other Asian background, Black/African/Caribbean/Black British, and Other ethnic group respectively.

Apart from Indian, Chinese, and other Asian background ethnic groups, all the other ethnic group's linear regression line closely follows the observed original data. Now, let us examine Table 2 and draw conclusions. The p-value is significantly lower than 0.05 for all the ethnic groups. Based on the p-values we can conclude that the predictor variables are statistically significant.

Considering the Beta estimate, we can observe from Table 2 that all the linear regression models have a positive beta estimate which suggests that the models have an increasing slope which indeed suggests that there is a positive tendency for the line to increase linearly with respect to independent variable which in this case is 'year'.

By considering the F-statistic, let us analyse the relationship between predictor and response variables. From Table 2, we can observe that F-stat value for the ethnic groups is significantly higher than 1. So, it is evident that there is a relation between predictor and response variables.

Let us now examine whether the linear regression model fits the original given data by considering the observed multiple R squared value. Pakistani, Bangladeshi, and Black/African/Caribbean/Black British have R2 values greater than 90% which is closer to 100%. This indicates that the linear regression model fits perfectly for the original data and can be used for further prediction. Mixed/multiple ethnic background, Indian, Other Asian background and Other ethnic group have R2 values in between 80% and 90% which indicates that the models can be used for those respective ethnic groups too. But for Chinese ethnicity, the R2 value is 0.6338 or 63.38% which is farther from the optimal value of 1. Therefore, the simple linear regression model for Chinese ethnicity is not a perfect model.

In conclusion, we can say that mathematically, the simple linear regression model is a good fit for the observed data, for the internet user data, for the considered ethnic groups in this section except Chinese. The internet usage among these ethnicities tend to increase linearly over the succeeding years.

Prediction for the next 5 years:

- Mixed/multiple ethnic background

1	2	3	4	5
587.3929	615.6190	643.8452	672.0714	700.2976

- Indian

1	2	3	4	5
1169.607	1204.048	1238.488	1272.929	1307.369

- Pakistani

1	2	3	4	5
845.5357	889.4048	933.2738	977.1429	1021.0119

- Bangladeshi

1	2	3	4	5
338.5714	359.5595	380.5476	401.5357	422.5238

- Chinese

1	2	3	4	5
294.1786	302.8571	311.5357	320.2143	328.8929

- Other Asian background

1	2	3	4	5
598.6429	620.7857	642.9286	665.0714	687.2143

- Black/African/Caribbean/Black British

1	2	3	4	5
1455.000	1520.917	1586.833	1652.750	1718.667

- Other ethnic group

```
       1        2        3        4        5
840.8929 870.7857 900.6786 930.5714 960.4643
```

Analysis of Ethnic group data using Multiple Linear Regression:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2865.55  -190.41    -8.85   203.75  2512.55

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     -259407.59   70802.42  -3.664 0.000516 ***
Year                                               128.70      35.09   3.667 0.000511 ***
Ethnicity_grpBlack/African/Caribbean/Black British  914.25     341.15   2.680 0.009420 **
Ethnicity_grpChinese                                 11.00     341.15   0.032 0.974381
Ethnicity_grpIndian                                 770.50     341.15   2.259 0.027441 *
Ethnicity_grpMixed/multiple ethnic background       216.25     341.15   0.634 0.528490
Ethnicity_grpOther Asian background                 254.87     341.15   0.747 0.457827
Ethnicity_grpOther ethnic group                     462.25     341.15   1.355 0.180344
Ethnicity_grpPakistani                              404.00     341.15   1.184 0.240845
Ethnicity_grpwhite                                38617.87     341.15 113.198  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 682.3 on 62 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9968
F-statistic:  2485 on 9 and 62 DF,  p-value: < 2.2e-16
```

Fig2.10 Multiple linear regression parameters – Age group

Figure 2.10 represents the parameters calculated after executing the given set of data with ethnic group to the multiple linear regression.

The p-value is approximately 2.2×10^{-16} which is significantly lower than 0.05 value. Apart from the p- value, R2 value is approximately 0.9972 which is closer to the optimal value 1. F – statistic is 2485 which is significantly higher than 1. All these parameter values shows that the given data fits with multiple linear regression model when we consider the ethnic group as an independent variable.

ANALYSIS OF INTERNET AND INTERNET NON-USERS WITH RESPECT TO ECONOMIC ACTIVITY:

In this section, we analyse the effect of Economic activity on the usage of internet in the United Kingdom. The economic activity groups under consideration are Employee, Self-employed, Government employment & training programmes, Unpaid family worker, Unemployed, Student, Retired, and Inactive

Simple linear regression is applied to model each economic activity group with respect to their internet user count and it has been analysed separately before applying the multiple linear regression to analyse the ethnic group data for a compiled conclusive analysis.

Table 3 represents the parameters of linear regression model with respect to the internet user data with different economic activity groups over the years 2014 to 2021.

Age Group	P – Value	R2	F-Statistic	Beta estimate (β)
Employee	3.739e-07	0.9894	559.9	5.630e+02
Self-employed	0.0001867	0.9167	66.01	160.38
Government employment & training programmes	0.07244	0.4411	4.736	-7.119
Unpaid family worker	0.04336	0.5205	6.514	3.417
Unemployed	0.0003257	0.8999	53.95	-155.06
Student	0.8882	0.003573	0.02151	-0.631
Retired	2.847e-10	0.999	6184	4.606e+02
Inactive	4.509e-05	0.9479	109.2	133.36

Table 3 - Summary of parameters of Linear Regression model with respect to Economic Groups

Now let us plot the linear regression model graph for each ethnic group for internet user count against years.

Employee:

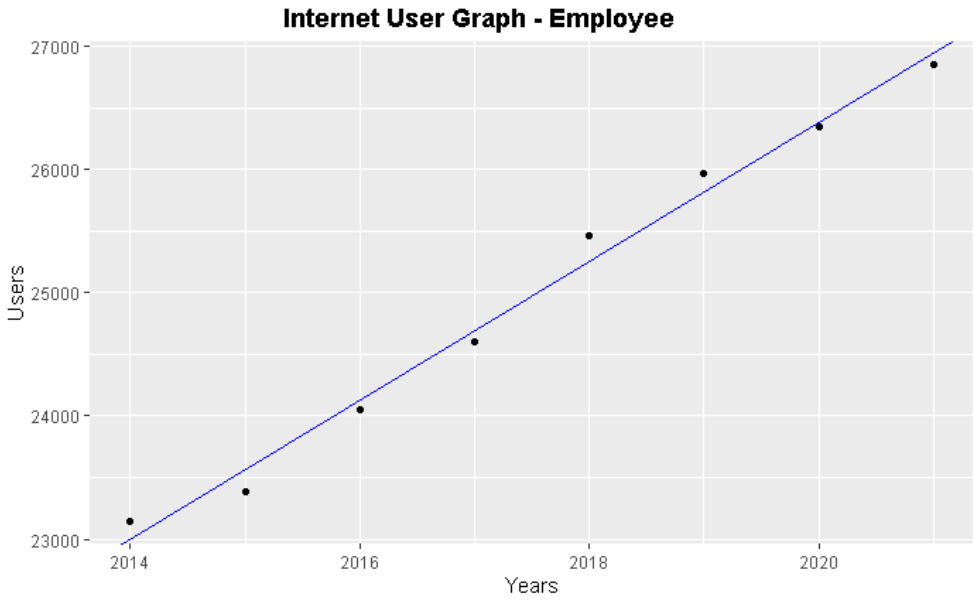

Fig 3.1 Internet User linear regression model Economic activity – Employee

Figure 3.1 represents the internet usage linear regression model graph for the Economic activity - Employee.

As we can observe from the graph, the linear regression is closely following the original data points. While interpreting the parameters from the mathematical calculations, we observed that the p-value 3.739×10^{-07} which is significantly lesser than 0.05. Based on the p-value we can conclude that the predictor variable (in this case economic activity - employee) is significant.

We observed that there is a positive beta estimate ($\beta = 563$) which indicates that with increase in years, the count of internet users also increases.

Now let us find out the relationship between predictor and response variables. In this model, F=559.9 which is significantly greater than 1. So, it can be concluded that there is a relationship between predictor and response variables.

By calculating the R2 (multiple R-squared) value, let us predict whether the model is fit or not. We observed the R2 value to be 0.9894 which is closer to 1. So, we can conclude that the model is fit to the original data.

Conclusion:

Both mathematically and visually, the simple linear regression model is a good fit for the observed data, for the internet user data, for the employee category. The internet usage among this economic group tends to increase linearly over the succeeding years.

Prediction for the next 5 years (Employee):

1	2	3	4	5
27511.21	28074.26	28637.31	29200.36	29763.40

The prediction for 5 years from 2022 to 2026 is represented above where '1' represents the year to '5' representing the year 2026. The values represent the internet user count for each year.

Let us summarize, the analysis for all the other economic activity groups by analysing the plots and taking Table 3 in consideration, in the following section.

Analysis for all other economic activity categories:

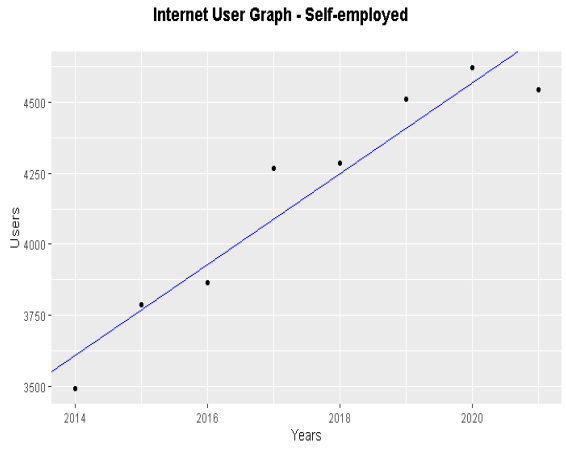

Fig 3.2 Internet User linear regression model Economic activity – Self employed

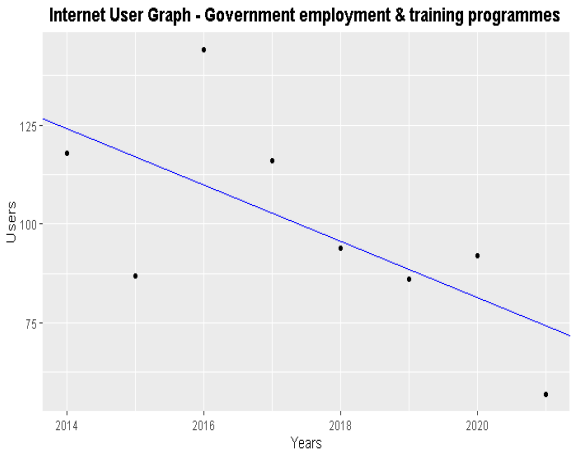

Fig 3.3 Internet User linear regression model Economic activity – Government employment & training programmes

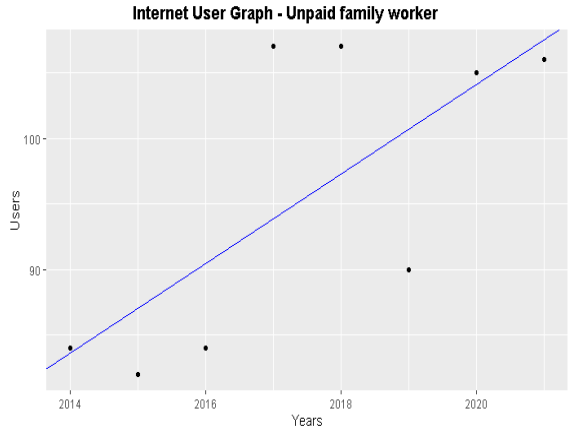

Fig 3.4 Internet User linear regression model Economic activity – Unpaid family worker

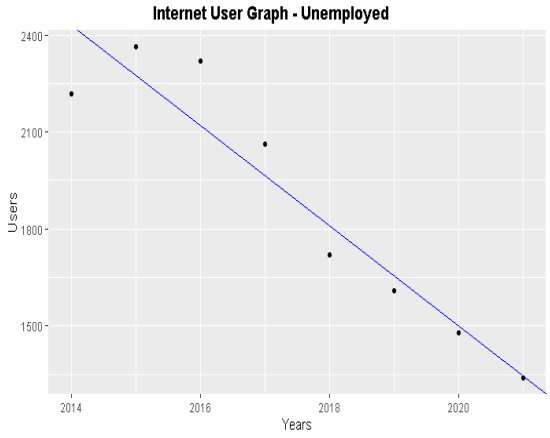

Fig 3.5 Internet User linear regression model Economic activity – Unemployed

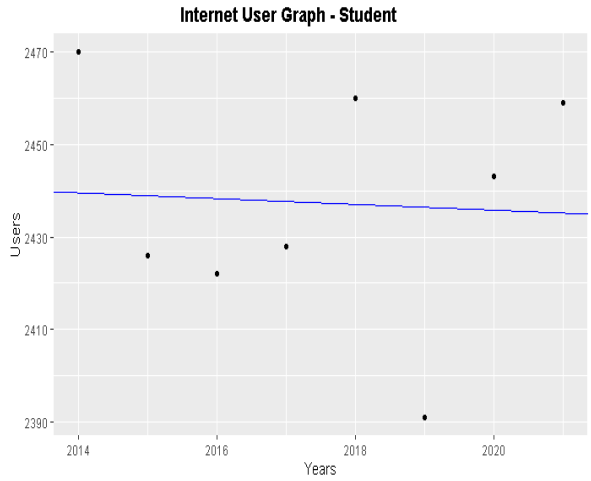

Fig 3.6 Internet User linear regression model Economic activity – Student

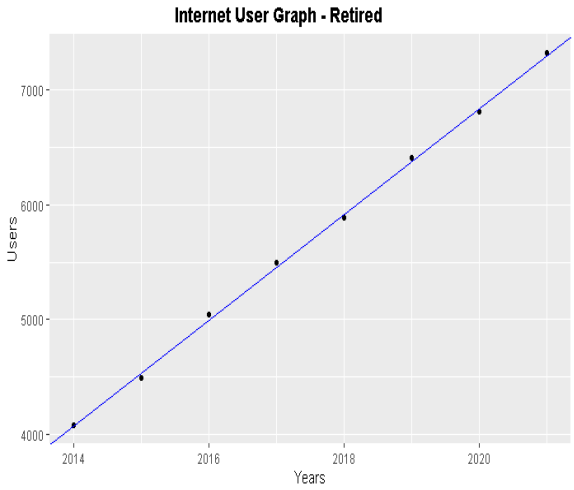

Fig 3.7 Internet User linear regression model Economic activity – Retired

Figure 3.2, 3.3, 3.4, 3.5, 3.6, 3.7 represents the internet usage linear regression model graph for the economic activity Employee, Self-employed, Government employment & training programmes, Unpaid family worker, Unemployed, Student, Retired, and Inactive respectively.

Apart Government employment & training programmes, Student, and Unpaid family worker, all the other economic activity group's linear regression line closely follows the observed original data. Now, let us examine Table 3 and draw conclusions. The p-value is significantly lower than 0.05 for employee, self-employed, unemployed, retired and inactive economic groups. P-values for Government employment & training programmes and student categories are higher than 0.05 which states that the predictor variables are statistically insignificant for those two particular activity groups. The p-value for unpaid family worker is 0.04336, which is lower than 0.05 but not significantly lower. Based on the p-values we can conclude that the predictor variables are statistically significant for employee, self-employed, unemployed, retired and inactive economic groups.

From Table 3, we can observe that there is a positive beta estimate (β) for employee, self-employed, unpaid family worker, retired, and inactive economic groups which indicates that with increase in years, the count of internet users also increases for these groups. But we can observe a negative beta estimate for Government employment & training programmes, Unemployed and Student categories which indicates that the number of internet users tend to decrease with time.

From Table 3, we can observe that F-stat is significantly higher than 1 for employee, self-employed, unemployed, retired, and inactive. So, it is evident that there is a relation between predictor and response variables for these groups. But the F-stat value is 0.02151 for 'student' which is lower than 1, which states that the relation between predictor and response variable is absent when this model is taken into consideration.

Except Government employment & training programmes, Student, and unpaid family worker categories, all the other groups have multiple R-squared value closer to 1. This proves that the simple linear model is fitting with the observed original data for all the economic activity groups except Government employment & training programmes, Student, and unpaid family worker categories.

Prediction for the next 5 years – Economic groups:

- Self employed

```
       1        2        3        4        5
4893.464 5053.845 5214.226 5374.607 5534.988
```

- Unemployed

```
       1        2        3        4        5
1190.6071 1035.5476  880.4881  725.4286  570.3690
```

- Retired

```
       1        2        3        4        5
7765.286 8225.905 8686.524 9147.143 9607.762
```

- Inactive

```
       1        2        3        4        5
5117.607 5250.964 5384.321 5517.679 5651.036
```

Government employment & training programmes, Student, and unpaid family worker categories are excluded from predicting the future internet user value because of the fact that the data from these economic group does not fit to the simple linear regression.

Analysis of Economic activity group data using Multiple Linear Regression:

```
Residuals:
    Min      1Q  Median      3Q     Max
-1323.9  -274.2    -7.5   283.5  1363.9

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     -267058.63   63723.78  -4.191 0.000102 ***
Year                                               144.75      31.59   4.583 2.68e-05 ***
economic_grpGovernment employment & training programmes  -24878.25     289.48 -85.940  < 2e-16 ***
economic_grpInactive                            -20460.00     289.48 -70.677  < 2e-16 ***
economic_grpRetired                             -19285.00     289.48 -66.618  < 2e-16 ***
economic_grpSelf-employed                       -20805.75     289.48 -71.872  < 2e-16 ***
economic_grpStudent                             -22540.12     289.48 -77.863  < 2e-16 ***
economic_grpUnemployed                          -23089.12     289.48 -79.759  < 2e-16 ***
economic_grpUnpaid family worker                -24881.88     289.48 -85.952  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 579 on 55 degrees of freedom
Multiple R-squared:  0.9951,    Adjusted R-squared:  0.9943
F-statistic:  1384 on 8 and 55 DF,  p-value: < 2.2e-16
```

Fig3.8 Multiple linear regression parameters – Economic activity

Figure 3.8 represents the parameters calculated after executing the given set of data with economic activity group to the multiple linear regression.

The p-value is approximately 2.2×10^{-16} which is significantly lower than 0.05 value. Apart from the p- value, R2 value is approximately 0.9951 which is closer to the optimal value 1. F – statistic is 1384 which is significantly higher than 1. All these parameter values shows that the given data fits with multiple linear regression model when we consider the economic activity category as an independent variable.

ANALYSIS OF INTERNET AND INTERNET NON-USERS WITH RESPECT TO GENDER:

In this section, we analyse the effect of gender on the usage of internet in the United Kingdom.

We will analyse and predict the trend of internet user count according to gender. We will consider the linear regression parameters to estimate whether the linear model is fit for the given data. For the analysis, we can classify the gender into men and women and do the mathematical calculations separately.

Men:

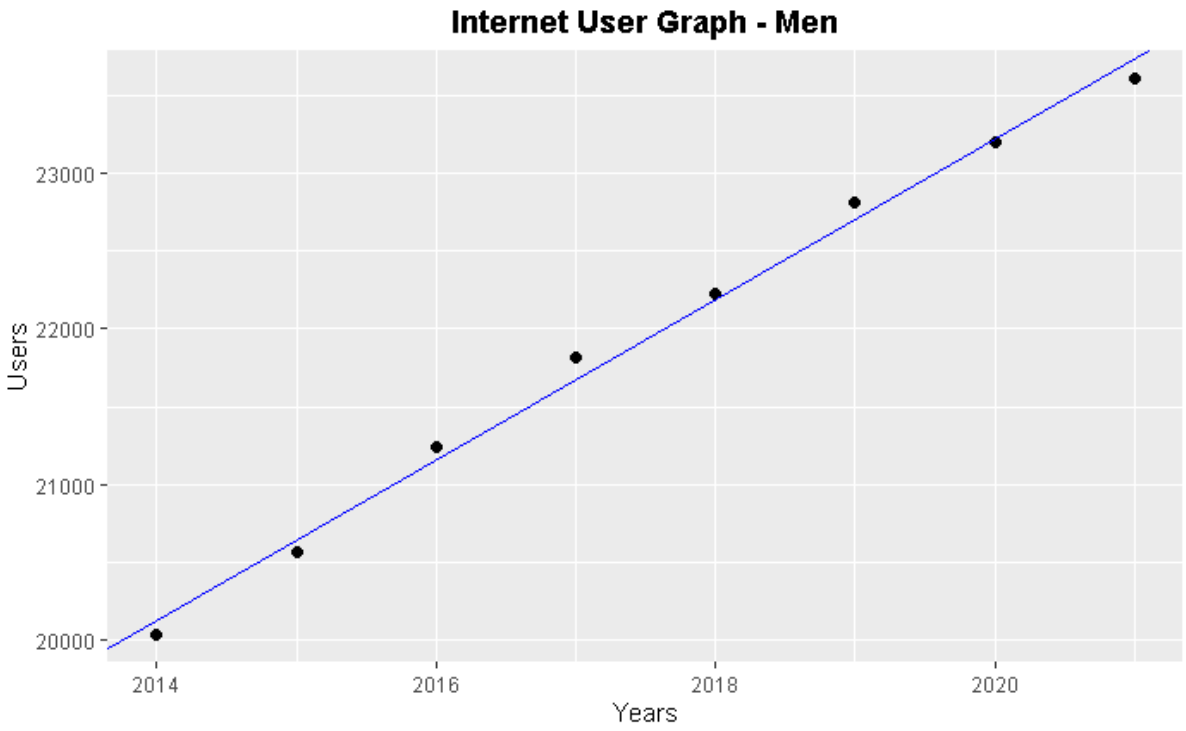

Fig 4.1 Internet User linear regression model – Men

Fig 4.1 represents the linear regression model of the internet user count for the gender Men from 2014 to 2021.

As we can observe from figure 4.1, the predicted linear regression model fits with the given data visually. Now, let us find out whether the model is fitting the original data statistically.

We observed that the p- value $7.82x\ 10^{-08}$ which is significantly lesser than 0.05. So, based on the p-value we can assume that the predictor is significant.

We observed that there is a positive beta estimate (β = 515.2) which indicates that with increase in years, the count of internet users also increases for men category.

Now let us find out the relationship between predictor and response variables. In this model, F=946.9 which is significantly greater than 1. So, it can be concluded that there is a relationship between predictor and response variables which is statistically significant.

By calculating the R2 (multiple R-squared) value, let us predict whether the model is fit or not. We observed the R2 value to be 0.9937 which is closer to 1. So, we can conclude that the model is fitting the original data.

By analysing the above parameters, we can conclude that the linear regression model is fitting perfectly with the given observed data.

Women:

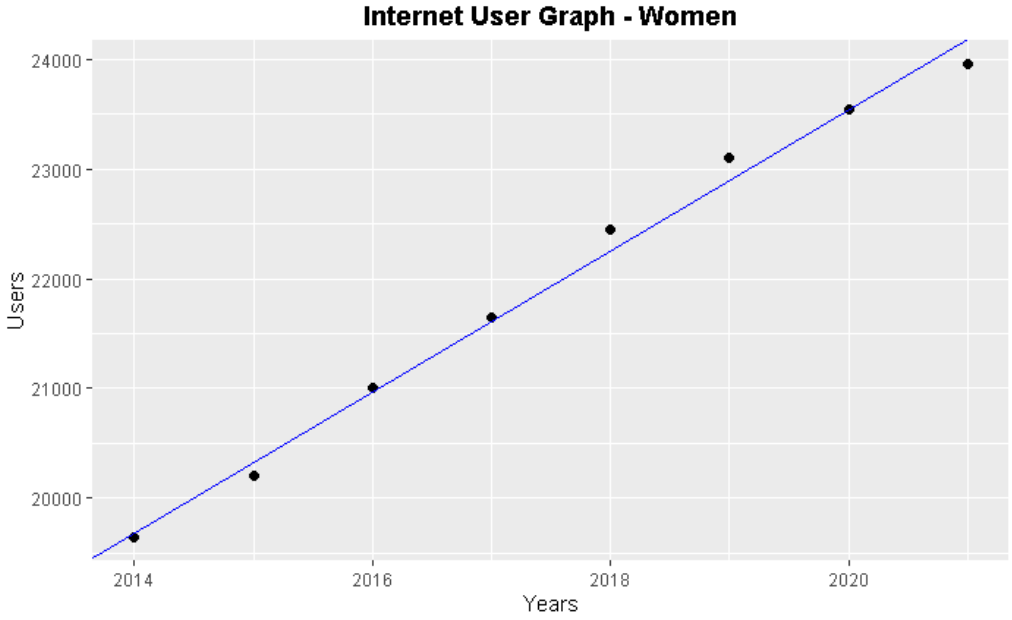

Fig 5.1 - Internet User linear regression model – Women

Fig 5.1 represents the linear regression model of the internet user count for the gender women from 2014 to 2021.

As we can observe from figure 5.1, the predicted linear regression model fits with the given data visually. Now, let us find out whether the model is fitting the original data statistically.

We observed that the p- value $1.956\ x\ 10^{-07}$ which is significantly lesser than 0.05. So, based on the p-value we can assume that the predictor is significant.

We observed that there is a positive beta estimate (β = 642.7) which indicates that with increase in years, the count of internet users also increases for women category.

In this model, F= 696.2 which is significantly greater than 1. So, it can be concluded that there is a relationship between predictor and response variables which is statistically significant.

We observed the R2 value to be 0.9915 which is closer to 1. So, we can conclude that the model is fitting the original data perfectly.

Analysis of Gender data using Multiple Linear Regression:

```
Residuals:
    Min      1Q  Median      3Q     Max
-358.46 -100.48    3.33  156.21  295.13

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.146e+06  4.575e+04 -25.050 2.18e-12 ***
Year         5.789e+02  2.268e+01  25.530 1.71e-12 ***
Genderwomen  3.250e+00  1.039e+02   0.031    0.976
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 207.8 on 13 degrees of freedom
Multiple R-squared:  0.9804,    Adjusted R-squared:  0.9774
F-statistic: 325.9 on 2 and 13 DF,  p-value: 7.822e-12
```

Fig5.2 Multiple linear regression parameters – Gender

Figure 5.2 represents the parameters calculated after executing the given set of data with gender to the multiple linear regression.

The p-value is 7.822×10^{-12} which is significantly lower than 0.05 value. Apart from the p- value, R2 value is approximately 0.9804 which is closer to the optimal value 1. F – statistic is 325.9 which is significantly higher than 1. All these parameter values shows that the given data fits with multiple linear regression model when we consider the gender group as an independent variable.

Conclusion:

By analysing the above parameters, we can conclude that the linear regression model is fitting perfectly with the given observed data for both men and women categories. Both categories showing a positive increase in linear regression line with respect to the independent variable. But as we cannot observe significant effect for both genders in changing or not changing the internet usage, we can say that the internet usage among people across the UK kingdom is independent of gender. The internet usage increases linearly irrespective of the gender factor.

Prediction for the next 5 years:

- Men

```
       1        2        3        4        5
24256.50 24771.67 25286.83 25802.00 26317.17
```

- Women

```
       1        2        3        4        5
24833.50 25476.17 26118.83 26761.50 27404.17
```

ANALYSIS OF INTERNET AND INTERNET NON-USERS WITH RESPECT TO DISABILITY:

In this section, we analyse and predict the internet user count for the years from 2017 to 2021 with respect to the disability. We separate the given data into Equality Act Disabled and Not Equality Act Disabled for the analysis.

Equality Act Disabled:

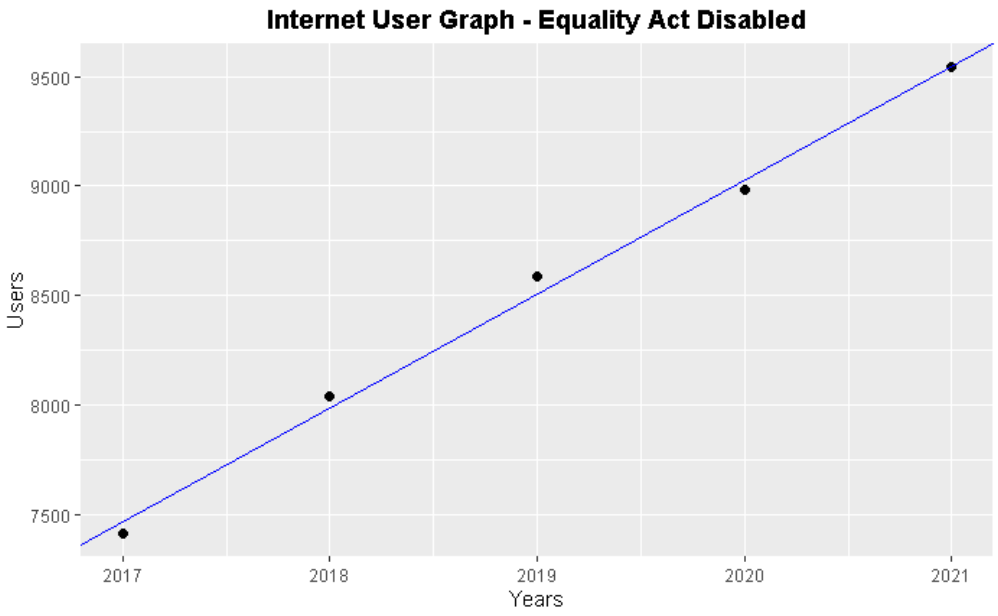

Fig 6.1 Internet User linear regression model – Disabled

Figure 6.1 represents the internet user simple linear regression model for the disabled population.

From the above plot, we can observe that the predicted regression line closely following the observed data with low variance. Let us observe the calculated p-value to derive the fit of the model. The p-value generated is approximately 0.0001475 which is lesser than 0.05 value when the 95% interval is considered. So the predictor is statistically significant. Also the beta estimate is a positive value (β = 520.6) which indicate that the linearly tends to increase with x-axis which is year.

The F statistic value is approximately equal to 604.5 which is much higher than 1, so the relationship between predictor and response variable is positive. Also, the calculated multiple R-squared value is 0.9951 which is much closer to optimal value 1. Therefore, the predicted model is perfectly fit for the given data.

So the mathematical calculations of the regression parameters gives evidence that the simple linear regression model is fitting the given set of data. The internet usage among the population of disabled people in the United Kingdom tend to increase over the years.

Not Equality Act Disabled:

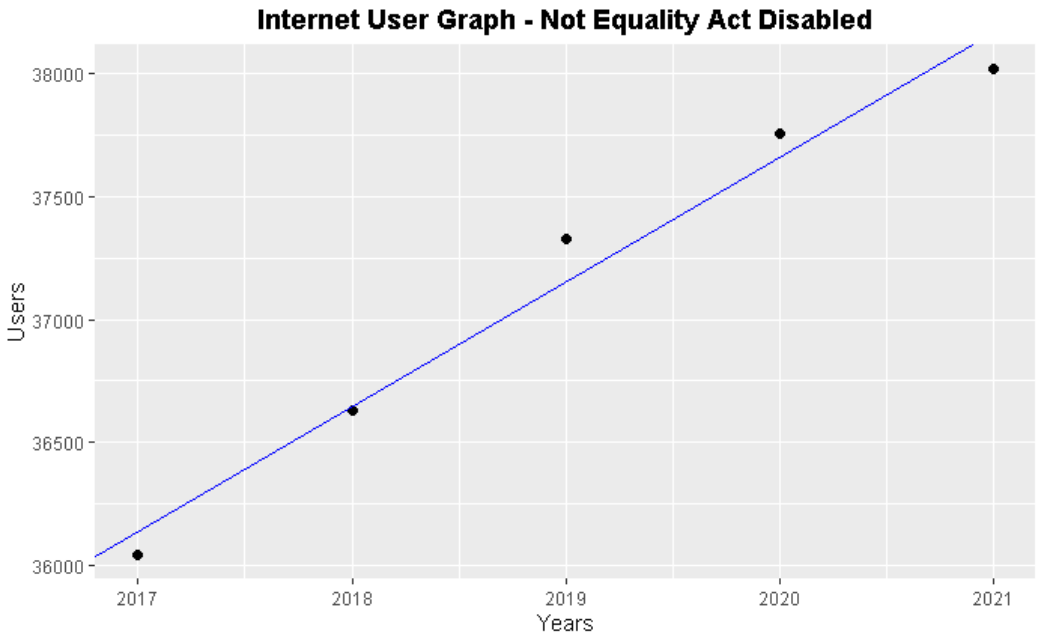

Fig 6.2 Internet User linear regression model – Not Disabled

Figure 6.2 represents the internet user simple linear regression model for the non - disabled population.

From figure 6.2, we can observe that the predicted regression line closely following the observed data with lesser variance. The p-value generated is approximately 0.001946 which is lesser than 0.05 value. So, the predictor is statistically significant. Also, the beta estimate is a positive value (β = 507.10) which indicate that the linearly tends to increase with x-axis which is year.

The F statistic value is approximately equal to 106.3 which is much higher than 1, so the relationship between predictor and response variable is positive. Also, the calculated multiple R-squared value is 0.9726 which is closer to the optimal value 1. Therefore, the predicted model is perfectly fit for the given data.

In conclusion, the mathematical calculations of the regression parameters give evidence that the simple linear regression model is fitting the given set of data. The internet usage among the population of non-disabled people in the United Kingdom tend to increase over the years.

Analysis of Disability group using Multiple Linear Regression:

```
Residuals:
     Min       1Q   Median       3Q      Max
-167.100  -64.738   -3.525   64.663  173.600

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.029e+06  5.031e+04  -20.45 1.68e-07 ***
Year                   5.139e+02  2.492e+01   20.62 1.58e-07 ***
Disabilitynot disabled 2.864e+04  7.048e+01  406.40  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111.4 on 7 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9999
F-statistic: 8.279e+04 on 2 and 7 DF,  p-value: 4.911e-16
```

Fig6.3 Multiple linear regression parameters – Disability group

Figure 6.3 represents the parameters calculated after executing the given set of data with disability group to the multiple linear regression.

The p-value is 4.911×10^{-16} which is significantly lower than 0.05 value. Apart from the p- value, R2 value is approximately 1. F – statistic is 8279 which is significantly higher than 1. All these parameter values shows that the given data fits with multiple linear regression model when we consider the disability group as an independent variable.

Conclusion:

Similar to gender classification, in the disability category too we can see the insignificance of the category with respect to the increase of people using internet in the UK. Irrespective of the disability factor, people tend to use internet and its increasing linearly year by year. But when we consider the linear growth of each category, we can observe a growth which is useful in implementing a marketing and sales strategy.

Prediction for next 5 years:

- Equality Act Disabled

```
       1        2        3        4        5
10074.6 10595.2 11115.8 11636.4 12157.0
```

- Not Equality Act Disabled

```
       1        2        3        4        5
38677.7 39184.8 39691.9 40199.0 40706.1
```

CONCLUSION

In this project, the analysis and prediction, of the internet usage in the UK across different categories which are age, sex, ethnicity, economic activity and disability, has been accomplished. Diving deep into each category, it was able to find out significant effects of each of these categories on the internet usage. From the analysis, one category was dominant to other categories in determining the trend of internet usage. It was the economic activity sector. More diverse results were derived when economic activity segment was analysed. It was visible that people who had a job or makes money tend to use more internet. There might be two reasons for this trend. One, more people are getting employed in UK or two, the internet rates are going up which makes it difficult for the less fortunate to continue using internet. But this factor does not affect the overall growth of internet users across the United Kingdom. In fact, it is easier for the advertising team to tunnel market to the population which has a job or makes money. One important observation which can be analysed in the economic activity sector is the linear growth of internet users among retired population. It can be considered a newly introduced population into the world of technology, so if the company could pierce into the 'retired' population through different marketing and sales strategies, it could be game changing.

Irrespective of age, people tend to use internet and it shows a growing trend among every age group except age group 16 – 24 which is indeed surprising. But this observation is negligible as the overall trend of internet users is positive. Gender and disability factor does not play a major role in altering the trend of internet usage. Both men and female internet usage tend to be increasing at a linear rate. It can be said the same disabled and non-disabled population. When the ethnicity of the population is considered, white ethnic group accounts for majority of the internet usage. This is expected as most of the population in the United Kingdom consist of people from White ethnicity. But as a positive note, the internet user count for all other ethnic groups tend to increase each year even though the count is incomparable to the White ethnicity. So, more focus can be given to the white ethnic group while marketing.

Overall it can be concluded that the scope is high for introducing the company to the telecommunication sector in the United Kingdom and it is feasible.

REFERENCES

- https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/
- https://towardsdatascience.com/simple-linear-regression-and-ols-introduction-to-the-theory-1b48f7c69867
- https://datascienceplus.com/linear-regression-with-healthcare-data-for-beginners-in-r/

APPENDIX

Code for Age group analysis:

Read the necessary packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Appuzzz/University/Sem 1/LSM/project
```

```
library(readxl)
```

Reading data of first workbook with internet users and non users based on age-group only.

```
workbook1 <- read_excel(here('internet_users.xlsx'))
```

```
workbook1_1 <-
  workbook1 %>%
  mutate(age_grp = factor(Age))

workbook1_2 <-
  workbook1_1 %>%
  filter(Age == 16)

workbook1_3 <-
  workbook1_1 %>%
  filter(Age == 25)

workbook1_4 <-
```

```
  workbook1_1 %>%
  filter(Age == 35)

workbook1_5 <-
  workbook1_1 %>%
  filter(Age == 45)

workbook1_6 <-
  workbook1_1 %>%
  filter(Age == 55)

workbook1_7 <-
  workbook1_1 %>%
  filter(Age == 65)

workbook1_8 <-
  workbook1_1 %>%
  filter(Age == 75)
```

Linear regression

```
m1_1 <- lm(internet_users_total ~ Year + age_grp, data = workbook1_1)
#m1_1 <- lm(internet_users_total ~ Year + Age, data = workbook1)
summary(m1_1)
```

```
par(mfrow=c(2,2))
plot(m1_1)
```

```
m1_16 <- lm(internet_users_total ~ Year , data = workbook1_2)
summary(m1_16)

predict(m1_16, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_25 <- lm(internet_users_total ~ Year, data = workbook1_3)
m1_35 <- lm(internet_users_total ~ Year, data = workbook1_4)
m1_45 <- lm(internet_users_total ~ Year, data = workbook1_5)
m1_55 <- lm(internet_users_total ~ Year, data = workbook1_6)
m1_65 <- lm(internet_users_total ~ Year, data = workbook1_7)
m1_75 <- lm(internet_users_total ~ Year, data = workbook1_8)
```

```
#anova(m1_1)
```

```
plot(workbook1_1$internet_users_total~workbook1_1$Year) + abline(m1_1)
```

```
## Warning in abline(m1_1): only using the first two of 8 regression coefficients
```

```
plot(workbook1_2$internet_users_total~workbook1_2$Year) + abline(m1_16)
```

```
ggplot(workbook1_2, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm") +
  labs(title="Internet User Graph for Age Group 16-24", x="Year", y="User Count")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
plot(workbook1_3$internet_users_total~workbook1_3$Year) + abline(m1_25)
```

```
ggplot(workbook1_3, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm")+
  labs(title="Internet User Graph for Age Group 25-34", x="Year", y="User Count")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
summary(m1_25)

predict(m1_25, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
plot(workbook1_4$internet_users_total~workbook1_4$Year) + abline(m1_35)
```

```
ggplot(workbook1_4, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm")+
  labs(title="Internet User Graph for Age Group 35-44", x="Year", y="User Count")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
summary(m1_35)

predict(m1_35, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
plot(workbook1_5$internet_users_total~workbook1_5$Year) + abline(m1_45)
```

```
ggplot(workbook1_5, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm")+
  labs(title="Internet User Graph for Age Group 45-54", x="Year", y="User Count")
```

```
summary(m1_45)

predict(m1_45, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
plot(workbook1_6$internet_users_total~workbook1_6$Year) + abline(m1_55)
```

```
ggplot(workbook1_6, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm")+
  labs(title="Internet User Graph for Age Group 55-64", x="Year", y="User Count")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```r
summary(m1_55)

predict(m1_55, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```r
plot(workbook1_7$internet_users_total~workbook1_7$Year) + abline(m1_65)
```

```r
ggplot(workbook1_7, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm")+
  labs(title="Internet User Graph for Age Group 65-74", x="Year", y="User Count")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```r
summary(m1_65)

predict(m1_65, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```r
plot(workbook1_8$internet_users_total~workbook1_8$Year) + abline(m1_75)
```

```r
ggplot(workbook1_8, aes(x = Year, y = internet_users_total)) +
            geom_point() +
  geom_smooth(method="lm")+
  labs(title="Internet User Graph for Age Group 75+", x="Year", y="User Count")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```r
summary(m1_75)

predict(m1_75, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

Code for Gender analysis:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Appuzzz/University/Sem 1/LSM/project
```

```
library(readxl)
```

Including Plots

You can also embed plots, for example:

```
gender_df <- read_excel(here('gender.xlsx'))
```

```
gender1 <-
  gender_df %>%
  mutate(Gender = factor(Gender))
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
gender_men <-
  gender1 %>%
  filter(Gender == "men")
```

```
gender_women <-
  gender1 %>%
  filter(Gender == "women")
```

```
li_gender <- lm(Users ~ Year + Gender, data = gender1)
#mi_1 <- lm(internet_users_total ~ Year + Age, data = workbook1)
summary(li_gender)
```

```
li_gender_men <- lm(Users ~ Year, data = gender_men)
summary(li_gender_men)

li_gender_women <- lm(Users ~ Year, data = gender_women)
summary(li_gender_women)

plot(gender_men$Users~gender_men$Year) + abline(li_gender_men)
```

```
plot(gender_women$Users~gender_women$Year) + abline(li_gender_women)
```

```
gg_men <- ggplot(gender_men, aes(x=Year ,y=Users)) + geom_point(size=2.2) + geom_abline(slope = coef(li
gg_men + ggtitle("Internet User Graph - Men") + xlab("Years") + ylab("Users") + theme(plot.title = elem
```

```
gg_women <- ggplot(gender_women, aes(x=Year ,y=Users)) + geom_point(size=2.2) + geom_abline(slope = coe
gg_women + ggtitle("Internet User Graph - Women") + xlab("Years") + ylab("Users") + theme(plot.title =
```

```
predict(li_gender_men, data.frame(Year=c(2022,2023,2024,2025,2026)))

predict(li_gender_women, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

Code for Disability factor analysis:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Appuzzz/University/Sem 1/LSM/project
```

```
library(readxl)
```

Including Plots

You can also embed plots, for example:

```
df_disability <- read_excel(here('disability.xlsx'))
```

```
dis_df1 <-
  df_disability %>%
  mutate(Disability = factor(Disability))
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
disabled_df <-
  dis_df1 %>%
  filter(Disability == "disabled")
```

```
not_disabled_df <-
  dis_df1 %>%
  filter(Disability == "not disabled")
```

```
d1 <- lm(Users ~ Year + Disability, data = dis_df1)
#m1_1 <- lm(internet_users_total ~ Year + Age, data = workbook1)
summary(d1)
```

```
d_dis <- lm(Users ~ Year, data = disabled_df)
summary(d_dis)



plot(disabled_df$Users~disabled_df$Year) + abline(d_dis)
```

```
gg_dis <- ggplot(disabled_df, aes(x=Year ,y=Users)) + geom_point(size=2.2) + geom_abline(slope = coef(d
gg_dis + ggtitle("Internet User Graph - Equality Act Disabled") + xlab("Years") + ylab("Users") + theme
```

```
predict(d_dis, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
d_nt_dis <- lm(Users ~ Year, data = not_disabled_df)
summary(d_nt_dis)



plot(not_disabled_df$Users~not_disabled_df$Year) + abline(d_nt_dis)
```

```
gg_not_dis <- ggplot(not_disabled_df, aes(x=Year ,y=Users)) + geom_point(size=2.2) + geom_abline(slope =
gg_not_dis + ggtitle("Internet User Graph - Not Equality Act Disabled") + xlab("Years") + ylab("Users")
```

```
predict(d_nt_dis, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

Code for ethnicity analysis:

```
library(tidyverse)
```

```
library(here)
```

```
## here() starts at C:/Appuzzz/University/Sem 1/LSM/project
```

```
library(readxl)
```

Including Plots

You can also embed plots, for example:

```
df <- read_excel(here('internet_users_ethnicity.xlsx'))
```

```
df1 <-
  df %>%
  mutate(Ethnicity_grp = factor(Ethnicity))
```

```r
df2 <-
  df1 %>%
  filter(Ethnicity_grp == "White")
```

```r
df3 <-
  df1 %>%
  filter(Ethnicity_grp == "Mixed/multiple ethnic background")

df4 <-
  df1 %>%
  filter(Ethnicity_grp == "Indian")

df5 <-
  df1 %>%
  filter(Ethnicity_grp == "Pakistani")

df6 <-
  df1 %>%
  filter(Ethnicity_grp == "Bangladeshi")

df7 <-
  df1 %>%
  filter(Ethnicity_grp == "Chinese")

df8 <-
  df1 %>%
  filter(Ethnicity_grp == "Other Asian background")

df9 <-
  df1 %>%
  filter(Ethnicity_grp == "Black/African/Caribbean/Black British")

df10 <-
  df1 %>%
  filter(Ethnicity_grp == "Other ethnic group")
```

```r
m1_eth <- lm(Users ~ Year + Ethnicity_grp, data = df1)
summary(m1_eth)
```

```r
m1_white <- lm(Users ~ Year, data = df2)
summary(m1_white)

plot(df2$Users~df2$Year) + abline(m1_white)
```

```
gg_white <- ggplot(df2, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_whit

gg_white + ggtitle("Internet User Graph - White") + xlab("Years") + ylab("Users") + theme(plot.title =
```

```
predict(m1_white, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_mixed <- lm(Users ~ Year, data = df3)
summary(m1_mixed)

plot(df3$Users-df3$Year) + abline(m1_mixed)
```

```
gg_mixed <- ggplot(df3, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_mixe

gg_mixed + ggtitle("Internet User Graph -  Mixed/multiple ethnic background
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```

```
predict(m1_mixed, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_indian <- lm(Users ~ Year, data = df4)
summary(m1_indian)

plot(df4$Users-df4$Year) + abline(m1_indian)
```

```
gg_indian <- ggplot(df4, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_ind

gg_indian + ggtitle("Internet User Graph -  Indian
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```


```
predict(m1_mixed, data.frame(Year=c(2022,2023,2024,2025,2026)))

m1_indian <- lm(Users ~ Year, data = df4)
summary(m1_indian)

plot(df4$Users-df4$Year) + abline(m1_indian)

gg_indian <- ggplot(df4, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_ind

gg_indian + ggtitle("Internet User Graph -  Indian
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```


```
predict(m1_indian, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_pakistani <- lm(Users ~ Year, data = df5)
summary(m1_pakistani)

plot(df5$Users-df5$Year) + abline(m1_pakistani)
```

```
gg_pak <- ggplot(df5, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_pakist

gg_pak + ggtitle("Internet User Graph -  Pakistani
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```

```
predict(m1_pakistani, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_ban <- lm(Users ~ Year, data = df6)
summary(m1_ban)

plot(df6$Users-df6$Year) + abline(m1_ban)
```

```
gg_ban <- ggplot(df6, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_ban)[[

gg_ban + ggtitle("Internet User Graph -  Bangladeshi
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```

```
predict(m1_ban, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_chinese <- lm(Users ~ Year, data = df7)
summary(m1_chinese)

plot(df7$Users-df7$Year) + abline(m1_chinese)
```

```
gg_chinese <- ggplot(df7, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_ch

gg_chinese + ggtitle("Internet User Graph -  Chinese
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```

```
predict(m1_chinese, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_asian <- lm(Users ~ Year, data = df8)
summary(m1_asian)
```

```
plot(df8$Users-df8$Year) + abline(m1_asian)
```

```
gg_asian <- ggplot(df8, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_asia

gg_asian + ggtitle("Internet User Graph -   Other Asian background") + xlab("Years") + ylab("Users") +
```

```
predict(m1_asian, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_black <- lm(Users ~ Year, data = df9)
summary(m1_black)

plot(df9$Users-df9$Year) + abline(m1_black)
```

```
gg_black <- ggplot(df9, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_blac

gg_black + ggtitle("Internet User Graph - Black/African/Caribbean/Black British
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```

```
predict(m1_black, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_other <- lm(Users ~ Year, data = df10)
summary(m1_other)

plot(df10$Users-df10$Year) + abline(m1_other)
```

```
gg_other <- ggplot(df10, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_oth

gg_other + ggtitle("Internet User Graph - Other ethnic group
") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.5"))
```

```
predict(m1_other, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Code of Economic activity analysis:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Appuzzz/University/Sem 1/LSM/project
```

```
library(readxl)
```

Including Plots

You can also embed plots, for example:

```
df <- read_excel(here('internet_users_economic.xlsx'))
```

```
df1 <-
  df %>%
  mutate(economic_grp = factor(eco_activity))
```

```r
df2 <-
  df1 %>%
  filter(economic_grp == "Employee")
```

```r
df3 <-
  df1 %>%
  filter(economic_grp == "Self-employed")

df4 <-
  df1 %>%
  filter(economic_grp == "Government employment & training programmes")

df5 <-
  df1 %>%
  filter(economic_grp == "Unpaid family worker")

df6 <-
  df1 %>%
  filter(economic_grp == "Unemployed")

df7 <-
  df1 %>%
  filter(economic_grp == "Student")

df8 <-
  df1 %>%
  filter(economic_grp == "Retired")

df9 <-
  df1 %>%
  filter(economic_grp == "Inactive")
```

```r
m1_eco <- lm(Users ~ Year + economic_grp, data = df1)
summary(m1_eco)
```

```r
m1_Employee <- lm(Users ~ Year, data = df2)
summary(m1_Employee)

plot(df2$Users~df2$Year) + abline(m1_Employee)
```

```r
gg_Employee <- ggplot(df2, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_E

gg_Employee + ggtitle("Internet User Graph - Employee ") + xlab("Years") + ylab("Users") + theme(plot.t
```

```r
predict(m1_Employee, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```r
m1_self <- lm(Users ~ Year, data = df3)
summary(m1_self)

plot(df3$Users~df3$Year) + abline(m1_self)
```

```r
gg_self <- ggplot(df3, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_self)

gg_self + ggtitle("Internet User Graph - Self-employed
 ") + xlab("Years") + ylab("Users") + theme(plot.title = element_text(face = "bold", hjust = "0.4"))

predict(m1_self, data.frame(Year=c(2022,2023,2024,2025,2026)))

m1_govt <- lm(Users ~ Year, data = df4)
summary(m1_govt)

plot(df4$Users-df4$Year) + abline(m1_govt)

gg_govt <- ggplot(df4, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_govt)

gg_govt + ggtitle("Internet User Graph - Government employment & training programmes") + xlab("Years")

predict(m1_govt, data.frame(Year=c(2022,2023,2024,2025,2026)))

m1_family <- lm(Users ~ Year, data = df5)
summary(m1_family)

plot(df5$Users-df5$Year) + abline(m1_family)

gg_govt <- ggplot(df4, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_govt)

gg_govt + ggtitle("Internet User Graph - Government employment & training programmes") + xlab("Years")

predict(m1_govt, data.frame(Year=c(2022,2023,2024,2025,2026)))

m1_family <- lm(Users ~ Year, data = df5)
summary(m1_family)

plot(df5$Users-df5$Year) + abline(m1_family)

gg_family <- ggplot(df5, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_fam

gg_family + ggtitle("Internet User Graph - Unpaid family worker") + xlab("Years") + ylab("Users") + ther

predict(m1_family, data.frame(Year=c(2022,2023,2024,2025,2026)))

m1_Unemployed <- lm(Users ~ Year, data = df6)
summary(m1_Unemployed)

plot(df6$Users-df6$Year) + abline(m1_Unemployed)

gg_unemployed <- ggplot(df6, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_

gg_unemployed + ggtitle("Internet User Graph - Unemployed") + xlab("Years") + ylab("Users") + theme(plo

predict(m1_Unemployed, data.frame(Year=c(2022,2023,2024,2025,2026)))

m1_Student <- lm(Users ~ Year, data = df7)
summary(m1_Student)

plot(df7$Users-df7$Year) + abline(m1_Student)

gg_student <- ggplot(df7, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_St

gg_student + ggtitle("Internet User Graph - Student") + xlab("Years") + ylab("Users") + theme(plot.titl
```

```
predict(m1_Student, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_Retired <- lm(Users ~ Year, data = df8)
summary(m1_Retired)

plot(df8$Users-df8$Year) + abline(m1_Retired)
```

```
gg_retired <- ggplot(df8, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_Re

gg_retired + ggtitle("Internet User Graph - Retired") + xlab("Years") + ylab("Users") + theme(plot.titl
```

```
predict(m1_Retired, data.frame(Year=c(2022,2023,2024,2025,2026)))
```

```
m1_Inactive <- lm(Users ~ Year, data = df9)
summary(m1_Inactive)

plot(df9$Users-df9$Year) + abline(m1_Inactive)
```

```
gg_inactive <- ggplot(df9, aes(x=Year ,y=Users)) + geom_point(size=1.5) + geom_abline(slope = coef(m1_I

gg_inactive + ggtitle("Internet User Graph - Inactive") + xlab("Years") + ylab("Users") + theme(plot.ti
```

```
predict(m1_Inactive, data.frame(Year=c(2022,2023,2024,2025,2026)))
```